The Debate over Inclusive Fitness as a Debate over Methodologies

Rubin, Hannah

Published in:
Philosophy of Science

DOI:
10.1086/694809

Publication date:
2018

The Debate over Inclusive Fitness as a Debate over Methodologies

Hannah Rubin*†
University of Groningen, University of Notre Dame

Abstract

This paper analyzes the recent debate surrounding inclusive fitness and argues that certain limitations ascribed to it by critics – such as requiring weak selection or providing dynamically insufficient models – are better thought of as limitations of the methodological framework most often used with inclusive fitness (quantitative genetics). In support of this, I show how inclusive fitness can be used with the replicator dynamics (of evolutionary game theory, a methodological framework preferred by inclusive fitness critics). I conclude that much of the debate is best understood as being about the orthogonal issue of using abstract versus idealized models.

1 Introduction

The mathematical framework of inclusive fitness was first introduced by Hamilton (1963, 1964) in order to help explain the evolution of social traits by kin selection and has helped to give new, intuitive explanations of a variety of traits including altruism, eusociality, parental care, and genomic imprinting (Grafen, 1984; Marshall, 2015, and references therein). In calculating inclusive fitness, one looks at the effects an organism has on other organisms' reproductive success, rather than just looking at the organism's own reproductive success. These effects are then weighted by the 'relatedness' of the organism to those organisms it affects.

In recent years, there has been an extensive debate surrounding inclusive fitness. Some authors argue that inclusive fitness calculations can be wrong (van Veelen, 2009), while others argue that it requires stringent assumptions and is less general than 'standard' natural selection (Nowak et al., 2010; Wilson, 2012; Allen et al., 2013). The response is that inclusive fitness calculations are not

*I would like to thank Simon Huttegger, Brian Skyrms, Jonathan Birch, Cailin OConnor, Kyle Stanford, Justin Bruner, three anonymous referees, members of the Social Dynamics Seminar at UC Irvine, and audiences at the Philosophy of ISHPSSB 2015 meeting for helpful comments.

†hannahmrubin@gmail.com

(merely by virtue of using the mathematical framework) susceptible to being wrong (Marshall, 2011) and do not require stringent assumptions like weak selection (Abbot et al., 2011; Marshall, 2015, etc.), additive payoffs (Queller, 1992; Taylor and Maciejewski, 2012; Birch, 2014b; Birch and Okasha, 2015, etc.), pairwise interactions (Taylor and Gardner, 2007; Abbot et al., 2011; Marshall, 2011, etc.), or special population structures (Taylor and Frank, 1996; Taylor and Gardner, 2007; Abbot et al., 2011; Taylor and Maciejewski, 2012; Marshall, 2015, etc.).

Critics of inclusive fitness often propose evolutionary game theory and/or population genetics as alternatives to the inclusive fitness framework (Traulsen, 2010; Nowak et al., 2010, 2011; Allen et al., 2013; Allen and Nowak, 2015). Often, the comparisons are made between very simple models in quantitative genetics, which abstract away from particular details of any given population, and more complex models arising out of population genetics, which often take into account more of the particular details. Here, we will look at how inclusive fitness can function in evolutionary game theory, which often makes idealizations rather than abstractions in order to achieve simple models. The difference between these two modeling strategies (using abstractions versus using idealizations) and how this relates to the inclusive fitness debate will be discussed more in sections 3.3 and 5. Looking at the way inclusive fitness can be incorporated into evolutionary game theory will help show where some of the disagreements about inclusive fitness arise and when inclusive fitness calculations might be expected to have the limitations ascribed to them by critics. It will also demonstrate how we can think of some parts of the debate as arising from different sides emphasizing different methodologies, rather than as disagreements over inclusive fitness as a way of calculating fitness.

First, I will introduce the framework of inclusive fitness and compare it to 'neighbor-modulated' fitness calculations in section 2. Then, in section 3, I will discuss the debate that has arisen around the inclusive fitness framework, focusing on issues which can be understood as arising from the different sides of the debate emphasizing different methodologies. In section 4, I will discuss how models using both neighbor-modulated and inclusive fitness are connected and provide a simple example to demonstrate these connections. Section 5 will provide a few ways to think about these connections and explain how they can help us understand some issues in the inclusive fitness debate. Finally, section 6 concludes.

2 Inclusive Fitness and Neighbor-Modulated Fitness

2.1 Basic Calculations

Inclusive fitness and the related concept of neighbor-modulated fitness were first proposed by Hamilton (1964). Roughly, the neighbor-modulated fitness of an organism is calculated by adding up the number of offspring the organism is

expected to have from some social interaction of interest. Inclusive fitness is an alternative mathematical framework in which fitness calculations track the offspring *caused by* a particular organism, rather than tracking the offspring an organism actually has. The offspring caused by the organism are then weighted according to a 'relatedness' parameter, which is a measure of how likely it is that the focal organism and its social partner share genetic material, relative to the rest of the population. The two types of fitness calculations provide alternative ways of partitioning the causal structure of social interactions. A more concrete description of the equations used in both frameworks will be provided below.

The inclusive fitness framework might initially seem counter-intuitive, so it is helpful to start with a basic observation: in general, a trait will increase in frequency when organisms with that trait have more offspring than the average organism in the population. To determine whether a trait of interest will increase in frequency, we want to see how many offspring organisms with that trait will have. Inclusive fitness gives us this information by telling us how many offspring are caused by an organism and how likely it is that these offspring are had by an organism with the trait of interest.

We can calculate inclusive fitness for a focal organism, i, by looking at the effects from all its social interactions relevant to our trait of interest. When i interacts with other organisms, it affects its own fitness by some amount (s_{ii}) and the fitness of another organism, j, by some amount (s_{ij}). The genotype of organism i also predicts, to a certain extent, the genotype of the social partner j. This relationship is described by r_{ij}. There will be more details on calculating r_{ij} in sections 2.2 and 4, but for now we can think of it as a measure of how likely it is that i and j share genetic material. We can then calculate inclusive fitness as follows:

$$f_i = \sum_j r_{ij} s_{ij} \tag{1}$$

This fitness calculation gives us information about how the population will evolve. It tells us how many offspring are had by organisms with the trait of interest, and since offspring tend to be like their parents, this gives us information about how the composition of a population is expected to change. Note that, although it is sometimes described this way, inclusive fitness is *not* calculated by counting the number of offspring an organism has and then adding all the offspring its relatives have (weighted by relatedness).

Compare the inclusive fitness approach to the neighbor-modulated fitness approach, where we look at an organism, i, and add up the effects of its social interactions on its own number of offspring. The neighbor-modulated fitness of organism i is then calculated as follows:

$$f_i = \sum_j s_{ji} \tag{2}$$

where s_{ji} is the effect i's social interaction with j has on i's fitness.[1] This gives us information about how many offspring i is expected to have and, since

[1]Note that the definition of neighbor-modulated fitness looks formally different from in-

i's offspring tend to be like i, about how the composition of a population is expected to change.[2]

2.2 Hamilton's Rule, the Price Equation, and Kin Selection

Hamilton's rule, famously associated with inclusive fitness, gives a condition for the increase of an altruistic behavior, where an organism performs an action that decreases its own fitness and increases the fitness of another. (An example of a model of the evolution of altruistic traits will be given in section 4.2.) It says simply that if the relatedness-weighted benefit of a trait exceeds its cost, then we should expect selection to favor that trait. That is, the trait is favored when:

$$bR - c > 0 \tag{3}$$

where b is the benefit to the focal organism's social partner and c is the cost to the focal organism.

Many results within the inclusive fitness framework, including Hamilton's rule, are derived from the Price equation, which is a general description of evolutionary change. Let f be the fitness of a trait in the population, relative to the average fitness in the population. Then, the Price equation describes expected evolutionary change in the following way:

$$\dot{E}(p) = Cov(f, p) \tag{4}$$

We can think of p as the average phenotypic value of the population, although p can actually represent anything a modeler might want to keep track of: phenotypic value, genetic value, frequency of a trait, etc. $\dot{E}(p)$ is then the change in the average value. The covariance term measures how fitness changes with differences in phenotype.[3]

When fitness effects are additive, that is, when the fitness effects on the recipient do not depend on the recipient's genotype/phenotype and fitness effects from all an organism's social interactions can simply be added up, we can derive equations for both inclusive fitness and neighbor-modulated fitness from the

clusive fitness as fitness effects are unweighted, while the fitness effects in inclusive fitness are weighted by a relatedness parameter. This apparent asymmetry disappears at the population level when we calculate the fitness of organisms with a certain trait. See section 4.1 for a calculation of neighbor-modulated fitness at the population level. For more information on the calculations of these two types of fitness, see (Frank, 1998, p. 48-9) and Birch (2016).

[2]Technically, both inclusive fitness and neighbor-modulated fitness include a baseline non-social fitness component, so these calculations are the fitness effects of the social trait of interest.

[3]There is sometimes a second term, $E_f(\dot{p})$, included which measures the fitness-weighted transmission bias, the difference between the phenotypic value of a parent and the average phenotypic value of their offspring. It is often assumed that $E_f(\dot{p}) = 0$, which is generally thought of as assuming there is no transmission bias. (Assuming that $E_f(\dot{p}) = 0$ is not exactly the same as assuming there is no transmission bias (van Veelen, 2005), but the details of what exactly it means to assume $E_f(\dot{p}) = 0$ are not crucial here.)

Price equation.[4] These equations are discussed further in the appendix, but here we will look at Hamilton's rule as derived from the Price equation. The (inclusive fitness version of) Hamilton's rule is:

$$\dot{E}(g) > 0 \text{ when } \beta_{s_{i-i}p} \cdot \frac{Cov(p, g')}{Cov(p, g)} - \beta_{s_{ii}p} > 0 \tag{5}$$

When we can interpret the covariance between an organism's phenotype and its own fitness ($\beta_{s_{ii}p}$) as a 'cost' and the covariance between an organism's phenotype and its social partner's fitness ($\beta_{s_{i-i}p}$) as a 'benefit', we have Hamilton's rule, where $R = \frac{Cov(p,g')}{Cov(p,g)}$. This measure of relatedness compares the covariance between a focal organism's phenotype, p, and its social partner's genotype, g', with the covariance between the focal organism's phenotype and its own genotype, g (Orlove and Wood, 1978). It is a measure of the degree to which the focal organism and its social partner are genetically related, or how likely it is that the fitness effects from a trait fall on organisms with the gene(s) encoding for the trait.

Section 4.1 and the appendix discuss how inclusive fitness results derived from the Price equation are related to the replicator dynamics, which is often used in game theoretic models, using methods drawn from Page and Nowak (2002). Section 4.2 will discuss how this definition of relatedness matches up with the definition of relatedness we will use in game theoretic models. Section 5.2 will discuss versions of Hamilton's rule which do not rely on the assumption of additive fitness components in relation to the results discussed here.

Relatedness is commonly thought of as a measure of the average kinship between interacting organisms when talking about kin selection for a trait. However, it is widely acknowledged that R, and many methods for calculating R, can be thought of as general measures of correlation between types (Marshall, 2015). In this case, R can measure how likely it is that altruists interact with other altruists regardless of whether that correlation is caused by interacting with kin or by some other mechanism, such as a green-beard effect where altruists are able to recognize and preferentially interact with other altruists.

Because inclusive fitness is often used in describing traits that evolve via kin selection, the terms 'inclusive fitness' and 'kin selection' are sometimes used interchangeably. However, it is important to distinguish inclusive fitness from kin selection. Inclusive fitness is a method of calculating fitness, as described above. Kin selection, on the other hand, refers to the selection of a trait due to benefits falling differentially on relatives. Inclusive fitness is a mathematical framework used to describe evolution of a trait; kin selection is a mechanism by which traits can evolve (Hamilton, 1975; Grafen, 2007a, among others).

Some of the critiques of inclusive fitness models are aimed at showing that kin selection has been less important as an evolutionary force than many inclusive

[4] The additivity of fitness effects requires satisfying these two conditions, which Birch (2016) refers to as *actor's control* and *weak additivity*. Actually, only the second condition is required to derive neighbor-modulated fitness while both are required to derive inclusive fitness. See Birch (2016) for a discussion of this.

fitness theorists presume (see Wilson, 2012, for example). Other parts of the criticism are aimed at the mathematical framework of inclusive fitness itself, such as claims that there are mathematical difficulties with the calculations in inclusive fitness (Nowak et al., 2010; Traulsen, 2010; Wilson, 2012). This paper will not discuss whether kin selection provides an adequate explanation of prosocial behavior. Instead, it looks at whether inclusive fitness can provide an adequate mathematical framework for use in evolutionary models. Kin selection is discussed only in considering how inclusive fitness can be used in models of traits evolving via kin selection. This focus will help us see which aspects of the debate are relevant to the inclusive fitness framework, and which pertain to kin selection explanations of the evolution of particular traits. Section 5 will discuss this further.

3 The Debate Surrounding Methods

There are several critiques levied against the inclusive fitness framework. This paper will address a couple of particularly important critiques which, as we will see, can be understood in light of an emphasis on different modeling techniques: the critiques that inclusive fitness requires the assumption of weak selection and cannot provide dynamically sufficient models. Here, I will give a description of these critiques and a brief motivation for thinking of them as arising from different sides of the debate emphasizing different methodologies. Section 5 provides a more detailed argument for this conclusion using material that will be laid out in section 4.

3.1 Weak Selection

First, inclusive fitness has been critiqued for requiring the assumption of weak selection. In assuming that there is weak selection, we assume that gene frequencies are not changing or that the changes in gene frequencies are small enough to be ignored.[5] This assumption is used in various ways in inclusive fitness models: in employing estimation methods for calculating relatedness, in ignoring higher-order effects or certain types of population structure, etc.

It is easy to see why certain methods of estimating relatedness require weak selection. For example, unless very special conditions hold, estimating relatedness using pedigrees, or family trees, requires that selection is weak. If gene frequencies are systematically changing in the population, the relatedness of an organism to its siblings, for example, will change as the genetic composition of its siblings changes (Grafen, 1984). However, calculating relatedness does not,

[5] One way to achieve this in a model is to write down fitness as the sum of two components: $f = f_0 + \delta f_x$. One of these, f_0, is the 'background' fitness, the fitness organisms get from things that are not related to the trait of interest. This is the same for all organisms. The fitness the organisms get from things related to the trait of interest, f_x, is then weighted by a parameter δ and as we take δ to zero, we approach the limit of weak selection. This is what Wild and Traulsen (2007) refer to as 'δ-weak selection'.

in general, require weak selection, and we can calculate how relatedness changes as gene frequencies change (Grafen, 1985; Birch, 2014a; Marshall, 2015).

The assumption of weak selection is also used because it allows one to ignore non-additive fitness effects. That is, the assumption of weak selection has been used to ignore things like synergistic effects (where organisms receive additional benefits from cooperation if they both cooperate) or the the effects of competition over resources. This is perhaps the more important use of the assumption of weak selection, as it allows one to separate the way an organism affects its own fitness (a self-effect) from the way it affects its social partner's fitness (an other-effect) in cases where the simplifying assumption of additive fitness components is false. Note that this critique also applies to neighbor-modulated fitness, as the fitness effects are similarly separated into components for self- and other-effect components. At some points in the debate, it seems that critics argue against the use of inclusive fitness (and neighbor-modulated fitness) because it requires weak selection in order to achieve the separation of fitness components. That is, without the assumption of weak selection, one is restricted to a special case in which fitness effects are additive, leading to the conclusion that inclusive fitness is less general than 'standard' natural selection (Nowak et al., 2010).

However, at some points it seems that critics want to claim that, whether or not fitness effects can be split into additive components, inclusive fitness calculations require weak selection. For instance, Nowak et al. (2010) claim that "...inclusive fitness theory cannot even be defined for non-vanishing selection; thus the assumption of weak selection is automatic" (SI 14). It is this second, stronger, claim that will addressed here. In section 4, the claim will be shown to be clearly false using modeling techniques from evolutionary game theory, one of the preferred frameworks of critics of inclusive fitness. Section 5 will then discuss how, if we read the debate as a debate about inclusive fitness theory as a set of methods rather than inclusive fitness theory as a framework for calculating fitness, we can make sense of this claim.

3.2 Dynamic Sufficiency

Inclusive fitness has also been criticized for not being able to provide dynamically sufficient models (Nowak et al., 2010; Wilson, 2012). In a dynamically sufficient model, information about the population at any particular time is enough to make predictions about the population at all future times. So, information about a population at some starting time is enough to be able to predict how the population will evolve at all future times. In a dynamically sufficient model, one can predict whether the population will reach an equilibrium, a state in which the population is no longer evolving, and what the population composition will be at the equilibrium should it reach one.[6] Critics of inclusive fitness argue

[6]This paper only deals with deterministic models, but stochastic models can also be dynamically sufficient. A stochastic model is dynamically sufficient when the information about the probability distribution over types at some starting time is enough to predict how the probability distribution will evolve at all future times and to predict the limiting distribution.

that it cannot be used to describe the evolutionary trajectories or end points of evolution (Nowak et al., 2010, SI4).

One reason this criticism might be leveled against inclusive fitness is the general reliance on the Price equation, which is not dynamically sufficient.[7] More specifically, the Price equation itself is neither dynamically sufficient or insufficient (because it merely expresses a mathematical identity), but it can be either depending on what sort of model it is used with. When we do have a dynamically sufficient model, the Price equation will correctly describe evolutionary change in the model, but will not itself give any additional predictions (van Veelen et al., 2012).

Because many of the results in inclusive fitness theory, like Hamilton's rule, are formulated in absence of a particular model, and because the focus is often on estimating the covariances rather than calculating them from an evolutionary model, we might not always get dynamically sufficient models within the framework. These estimations of parameters will only predict the evolutionary outcome if they do not change over time, which is not the case when selection is frequency dependent (Nowak et al., 2010; Allen et al., 2013). However, as we will see in section 4, the regression methods often emphasized in inclusive fitness theory are intimately connected with the sort of dynamically sufficient models preferred by critics of inclusive fitness.

3.3 The Debate Over Methodologies

Critics of inclusive fitness often propose population genetics or evolutionary game theory as alternative frameworks in which one can provide models that are dynamically sufficient and that do not require stringent assumptions like weak selection (Traulsen, 2010; Nowak et al., 2010, 2011; Allen et al., 2013; Allen and Nowak, 2015). It is not immediately clear how we should read this proposal, because although it is true that inclusive fitness tends to be used in quantitative genetics models (Frank, 2013) and is seen as primarily a quantitative method in spirit (Queller, 1992), it has been used in both game theoretic (Skyrms, 2002; van Veelen, 2009, 2011, etc.) and population genetics models (Rousset, 2002; Grafen, 2007b; Lehmann and Rousset, 2014, etc.). In fact, when Hamilton (1964) first proposed using inclusive fitness, he did so in the context of a population genetic model.

The methods used in quantitative genetics are designed to handle continuously varying traits, such as height or weight. In models of social behavior, a continuously varying trait could be the probability of performing an altruistic action. Models within quantitative genetics tend to emphasize simplicity and measurability. These models usually start with observations about phenotypes, or other easily measurable quantities, with few assumptions about the underlying genetics of a trait. This method of modeling involves *abstractions*, ignoring complicating details of the situation by merely leaving them out while

[7] Another reason, which will be discussed further in sections 3.3 and 5.2, is that many of the results which do not rely on the Price equation are focused solely on equilibrium analysis. See, for example, Taylor and Frank (1996).

still giving a description that is literally true (Godfrey-Smith, 2009). The Price equation is often used within this approach. As mentioned in section 2.2, many of the common results within inclusive fitness theory are derived from the Price equation.

By contrast, challenges to the inclusive fitness framework tend to come from population genetics (Frank, 2013, p. 1153). This is an approach that tends to start with specific assumptions (such as assuming we know the underlying genetics of a trait, the mutation rates, etc.), and make predictions based on these assumptions. Models within this approach tend to be dynamically sufficient, meaning that information about the population at any particular time is enough to make predictions about the population at all future times. The use of simplifying assumptions also means that these models make use of *idealizations* rather than abstractions. That is, they talk about populations which have features we know real populations do not have (e.g. infinite population size, no mutations, etc.) in order to provide a simple model. One way to think about models using idealizations is that they describe non-actual, fictional populations that we take to be similar to real populations in important ways (Godfrey-Smith, 2009). As mentioned, critics propose evolutionary game theory as an alternative to the inclusive fitness framework.[8] The replicator dynamics is often used within this approach. This dynamics requires many idealizing assumptions, which will be discussed in section 4.1.

The rest of this paper will look more closely at the use of inclusive fitness in evolutionary game theory, focusing on the replicator dynamics. Since inclusive fitness is not as commonly used in evolutionary game theory, this will help us see the benefits and drawbacks of using inclusive fitness in highly idealized models. This paper will also compare how inclusive fitness calculations can be used in evolutionary game theory with some of their uses in quantitative genetics. This comparison between the use of inclusive fitness within these two traditions for studying evolution will be helpful in understanding key issues in the debate, since they represent extremes of methodologies using idealizations and abstractions: the replicator dynamics of evolutionary game theory is highly idealized, while the Price equation often employed in quantitative genetics uses only abstractions. We will see how some of the disagreement arises out of the sides of the debate emphasizing different methodologies and how this relates to arguments over the usefulness of Hamilton's rule.

It is important to note that, while this distinction between abstract models in quantitative genetics and idealized models in evolutionary game theory is illuminating for the present purposes, it does not capture the full variety of modeling techniques within the two methodological traditions. There are evolutionary game theoretic models which make the assumption of weak selection in order to abstract away from genetic details and fail to be dynamically sufficient. For instance, Taylor and Frank (1996) employ a weak selection assumption, al-

[8] Evolutionary game theory and population genetics are sometimes seen as having distinct methods and other times they are seen as more or less continuous (Hammerstein and Selten, 1994, p. 953). They are loosely grouped together here because they are similar in that models within both approaches tend to start which specific assumptions and be dynamically sufficient.

lowing them to approximate regression coefficients using partial derivatives, in order to use standard maximization techniques for finding evolutionary stable strategies (p. 28). This method can be used to derive 'approximate' versions of Hamilton's rule, which will be described further in section 5.2.

This paper will focus on the special case where fitness effects are additive. This is a starting point to examine how inclusive fitness can be calculated in idealized evolutionary game theoretic models and to see if there is any benefit to using inclusive fitness in this context. We will see that the assumption of weak selection is not essential to the calculation of inclusive fitness and that one can build dynamically sufficient models using inclusive fitness. There is, of course, further work to be done to see whether and how this can extend into the more complicated cases generally talked about in inclusive fitness theory. The relationship between these results and general versions of Hamilton's rule, which do not require weak selection and do not assume additive fitness components, will be discussed in section 5.2. Note, however, that while the special case of additive fitness effects will not be applicable to many traits of interest in the real world, it is an important special case which has been studied extensively in a variety of contexts even outside of the inclusive fitness framework (Eliashberg and Winkler, 1981; Chakraborty and Harbaugh, 2007; Maciejewski et al., 2014, among others).

4 Inclusive Fitness in Evolutionary Game Theory

Inclusive fitness and neighbor-modulated fitness are commonly viewed as 'formally equivalent' in that they yield the same predictions in terms of the direction of evolutionary change. That is, they give the same conditions for when a social trait is favored by evolution (see Birch, 2016, and references therein). This section will show that, in the special case discussed above, we can prove further that they also give the same prediction for magnitude of evolutionary change. Section 4.1 will prove that the two calculations of fitness are equivalent when used with the replicator dynamics, a standard model from evolutionary game theory. These results are then compared to more common calculations of inclusive fitness in the appendix, which proves the equivalence between the replicator dynamics and both the neighbor-modulated and inclusive fitness calculations derived from the Price equation. Then, section 4.2 provides a simple example to illustrate the connections between these fitness calculations.

4.1 Inclusive Fitness and Neighbor-Modulated Fitness in Evolutionary Game Theory

In evolutionary game theoretic models, the replicator dynamics is a standard model of the evolutionary process. Under this dynamic, if the fitness of a trait is greater than the average fitness of the population, the frequency of the trait

will increase. The traits of interest dictate behavior in some social interaction, so a trait's fitness is determined by how well it does against the other possible traits in the population (in addition to the population composition). If x_t is the frequency of the trait of interest, and $f_t(x)$ its fitness in a population of composition x, the replicator dynamics is governed by the following equation:

$$\dot{x}_t = x_t[f_t(x) - \bar{f}(x)] \tag{6}$$

where $\bar{f}(x)$ is the average fitness in the population. There are a number of assumptions involved in using the replicator dynamics, notably that the population size is infinite and there are a finite number of traits.

Since we are trying to see whether the trait of interest is favored, we can calculate the fitness of organisms which have the trait and the fitness of those that do not in order to have a full description of evolutionary change according to the replicator dynamics. As mentioned, we will look at the case where there are additive fitness effects. If we assume further that there are pairwise interactions, we can denote organism i's social partner as $-i$. In this case, we can write the neighbor-modulated fitness of the organisms with the trait of interest as

$$\begin{aligned} f_t(x) &= P(T_{-i}|T_i) \cdot (s_{ii} + s_{i-i}) + P(N_{-i}|T_i) \cdot s_{ii} \\ &= s_{ii} + P(T_{-i}|T_i) \cdot s_{i-i} \end{aligned} \tag{7}$$

where $P(T_{-i}|T_i)$ is the probability an organism with the trait will interact with another organism that has the trait and where $P(N_{-i}|T_i)$ is the probability an organism with the trait will interact with an organism that does not have the trait. Similarly, the neighbor modulated fitness of organisms without the trait of interest is

$$f_n(x) = P(T_{-i}|N_i) \cdot s_{i-i} \tag{8}$$

where $P(T_{-i}|N_i)$ is the probability an organism that does not have the trait will interact with another organism that does have the trait.

The inclusive fitness of organisms with the trait of interest is (now using w for inclusive fitness to distinguish it from neighbor-modulated fitness, f)

$$w_t(x) = s_{ii} + Rs_{i-i} \tag{9}$$

and the inclusive fitness of not having the trait is 0. The relatedness between interacting organisms, R, is defined as a difference in conditional probabilities (Skyrms, 2002; van Veelen, 2009; Okasha and Martens, 2016). The relatedness of a focal organism to its social partner is the probability the social partner has a trait given the focal organism does, minus the probability the social partner has the trait given the focal organism does not:

$$R = P(T_{-i}|T_i) - P(T_{-i}|N_i) \tag{10}$$

This is a measure of the degree to which the focal organism's phenotype predicts

it's social partner's phenotype.[9] Since genotypes (to a certain extent) predict phenotypes, this can also be thought of as a measure of genetic relatedness.[10]

If we start with the replicator dynamics with neighbor-modulated fitness as our measure of fitness, we can show that it is equivalent to using the replicator dynamics with inclusive fitness as our measure of fitness:

$$\begin{aligned}
\dot{x}_t &= x_t[f_t(x) - \bar{f}(x)] \\
&= x_t[s_{ii} + P(T_{-i}|T_i) \cdot s_{i-i} - x_t(s_{ii} + P(T_{-i}|T_i) \cdot s_{i-i}) - x_n(P(T_{-i}|N_i) \cdot s_{i-i})] \\
&= x_t[s_{ii} - x_t s_{ii} + P(T_{-i}|T_i) \cdot s_{i-i} - x_t P(T_{-i}|T_i) \cdot s_{i-i} - (1 - x_t)P(T_{-i}|N_i) \cdot s_{i-i})] \\
&= x_t[s_{ii} + (P(T_{-i}|T_i) - P(T_{-i}|N_i))s_{i-i} - x_t s_{ii} - x_t(P(T_{-i}|T_i) - P(T_{-i}|N_i))s_{i-i}] \\
&= x_t[s_{ii} + Rs_{i-i} - x_t(s_{ii} + Rs_{i-i})] \\
&= x_t[w_t(x) - \bar{w}(x)]
\end{aligned}$$

That is, neighbor-modulated fitness and inclusive fitness are equivalent when used with the replicator dynamics, a standard model of evolution used in evolutionary game theory.[11]

The appendix shows further that, given the assumptions stated above, using the replicator dynamics is equivalent to the Price equation with either method of calculating fitness. That is, the following are equivalent descriptions of evolutionary change:

1. The replicator dynamics used with neighbor-modulated fitness
2. The replicator dynamics used with inclusive fitness
3. The Price equation used with neighbor-modulated fitness
4. The Price equation used with inclusive fitness

The equivalence between (1) and (3) is demonstrated in appendix A. The general strategy is the same as the one used in Page and Nowak (2002). First, show that the Price equation used with neighbor-modulated fitness (3) is descriptive of a population evolving according to the replicator dynamics used with neighbor-modulated fitness (1), then show that when there are a finite number of types (3) is also descriptive of a population evolving according (1). Using

[9] Why this is the right definition to use is shown in (Skyrms, 2002). For a demonstration that the assortment rate from Grafen (1979) commonly used in the replicator dynamics is equivalent to a covariance definition of relatedness derived from the Price equation, see (Marshall, 2015, chapter 5, note 1).

[10] Note that relatedness is *not* just the probability that the two organisms share the allele of interest. It is a measure of their genetic similarity relative to the genetic composition of the population as a whole. This is important because in studying altruism, for example, we want to know whether the benefits of altruistic acts fall on altruists sufficiently *more often* than they fall on non-altruists. That is, the benefits must fall on altruists rather than non-altruists with sufficient frequency to give them a reproductive advantage over non-altruists. We will see an example of how R depends on the population's genetic composition in section 4.2.

[11] For a discussion of the relationship between inclusive fitness and neighbor-modulated fitness in games that do not assume pairwise interactions, but with a constant relatedness, see van Veelen (2011).

the same strategy, we can show that (2) and (4) are equivalent. This is done in appendix B. Note that these four ways of modeling evolutionary change are shown to be equivalent in that that they give the same prediction for both the direction and magnitude of evolutionary change. This goes beyond what is commonly meant by the claim that neighbor-modulated fitness and inclusive fitness are equivalent, which is that they give the same prediction for the direction of evolutionary change (see Birch, 2016, and references therein).

The next section provides a simple model using inclusive fitness in the context of evolutionary game theory. This simple illustrative example will let us see, in more concrete terms, the benefits and disadvantages of using inclusive fitness in such an idealized setting. Section 5 discusses how to understand these equivalences in the context of the inclusive fitness debate.

4.2 A Simple Model: Altruism with Haploid Siblings

This section will provide an idealized model using haploid siblings to show how one can dynamically model relatedness within the inclusive fitness framework when selection is not weak. We will assume that these organisms either have the altruistic trait or not, which is completely determined by whether or not they receive a certain gene from their parent. So that the relationship between this model and Hamilton's rule is clear, we will assume that when an organism has the altruistic trait, it pays a cost c and bestows a benefit b on its social partner. When an organism lacks the altruistic trait, it does not pay the cost and does not benefit its social partner. In this model, an organism's social partner is its sibling. Based on these assumptions, we can calculate the inclusive fitness of altruists to be:

$$f_a = -c + Rb \tag{11}$$

The inclusive fitness of non-altruists is 0 because they do not perform any action (relevant to our trait of interest) that affects their own or their social partner's reproduction. Thus altruism will spread when $bR - c > 0$.

Since the relatedness of haploid siblings is determined by the genetic material they receive from their common parent, we can let p be the frequency of altruists in the parent generation and use this to calculate relatedness among the offspring. We will also account for a small mutation rate μ in the calculation of relatedness. Once we rewrite the probabilities (according to the definition of conditional probability) so that they are easier to calculate from the assumptions of the model, we can calculate the relatedness of an altruist to its haploid sibling in the following way:

$$\begin{aligned} R &= P(A_{-i}|A_i) - P(A_{-i}|N_i) \\ &= \frac{P(A_{-i}\&A_i)}{P(A_i)} - \frac{P(A_{-i}\&N_i)}{P(N_i)} \\ &= \frac{p(1-\mu)^2 + (1-p)\mu^2}{p(1-\mu) + (1-p)\mu} - \frac{p(1-\mu)\mu + (1-p)(1-\mu)\mu}{p\mu + (1-p)(1-\mu)} \end{aligned}$$

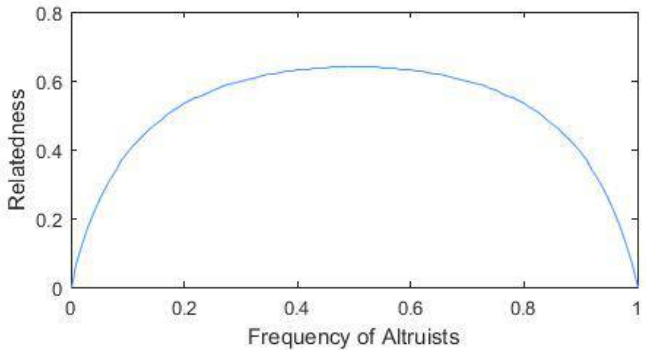

Figure 1: Relatedness graphed over the frequency of altruists in the parent population, for $\mu = 0.1$.

Briefly, here is how to understand this calculation. The numerator of the first term is the probability of two haploid siblings both being altruists. Since there are two ways to get two altruistic offspring, we can calculate this as the probability the parent is an altruist (p) times the probability it has two offspring without mutations ($(1-\mu)^2$), plus the probability the parent is a non-altruist ($1-p$) times the probability it has two offspring which both have a mutation (μ^2). The denominator of the first term is then the frequency of altruists in the offspring generation. These offspring can come from an altruist parent without mutation or from a non-altruist parent with mutation. The second term is calculated similarly. The numerator is the probability that a focal non-altruist will have an altruist sibling: the probability that the parent is an altruist and the focal organism mutates while its sibling does not plus the probability the parent is a non-altruist and the focal organism does not mutate while its sibling does. This is divided by the frequency of non-altruists in the offspring generation.

Figure 1 shows how R will change when the population's composition changes.[12] In particular, it shows that relatedness decreases as the population becomes more uniform.[13] To see why this is the case, it is easiest to look at the extremes of $p = 0$ and $p = 1$. When $p = 0$, the parent population is entirely composed of non-altruists. In the offspring generation, altruists only exist because of mutation. The probability an altruist has an altruist sibling is just μ, the probability that their sibling also has a mutation. However, the probability that a *non-altruist* has an altruist sibling is also μ, the probability that their sibling has a mutation. So $R = P(A_{-i}|A_i) - P(A_{-i}|N_i) = 0$. Similar reasoning applies when $p = 1$. The parent population is composed entirely of altruists, so any non-altruists in the offspring generation arise through mutation. This means that although altruists are likely to have altruist siblings, non-altruists

[12] This graph was created with a mutation rate of $\mu = 0.1$, which is a fairly high mutation rate. This mutation rate was chosen in order to make the graphs more readable. Results similar to those described in this section can be obtained with much smaller mutation rates.

[13] For a demonstration of this is a more general setting, see Rousset (2002).

are equally likely to have altruist siblings. So although $P(A_{-i}|A_i)$ is high at $1-\mu$, $P(A_{-i}|N_i)$ is also $1-\mu$, and $R=0$.

We can also calculate relatedness in this model using covariances or regressions. Since phenotypes in this idealized model are completely determined by genotypes (an organism with the altruistic gene is assumed to be an altruist), we can write:

$$R = \frac{Cov(p,g')}{Cov(p,g)} = \frac{Cov(g,g')}{Cov(g,g)} = \beta_{g'g} \tag{12}$$

For any population composition, we can perform a regression to calculate the value of R, and it will give the same value of relatedness as the probabilistic definition of relatedness. Figure 2 gives a way to visualize why this is the case. In this model, an organism's genetic value, g, is 1 if it has the gene for altruism and 0 if it does not. Thus there are four possible places for data points on a graph of g versus g': the four corners of the graph. Then, when we do a regression of g on g', what matters is how many data points are in each of these locations. When the focal organisms' genetic value is 1, its social partner's genotype will on average be $P(A_{-i}|A_i)$. Similarly, when the focal organisms' genetic value is 0, its social partner's genotype will on average be $P(A_{-i}|N_i)$. As shown in figure 2, this is the intercept of the regression, and the regression coefficient is $\beta_{g'g} = P(A_{-i}|A_i) - P(A_{-i}|N_i)$.

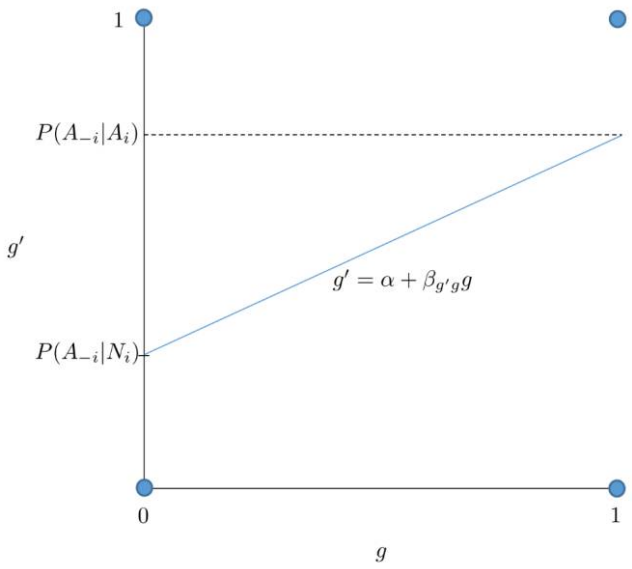

Figure 2: An illustration of $\alpha = P(A_{-i}|N_i)$ and $\beta_{g'g} = P(A_{-i}|A_i) - P(A_{-i}|N_i)$.

The inclusive fitness of altruists depends on R, so it also changes as the population composition changes. Figure 3 shows how the inclusive fitness of

altruists compares with the inclusive fitness of non-altruists over the possible population compositions, for $b = 18$ and $c = 10$. Since relatedness drops off as the population becomes uniform, the inclusive fitness of altruists drops off as the population becomes more uniform. For many possible values of b, c, and μ this means that altruists will have a fitness advantage for some area around $p = 0.5$, but their fitness will drop below the fitness of non-altruists as the population becomes more uniform.

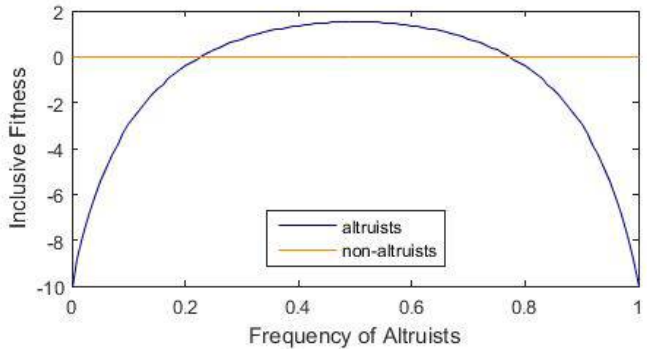

Figure 3: Inclusive fitness graphed over the frequency of altruists in the parent population, for $\mu = 0.1$, $b = 18$ and $c = 10$.

These calculations of relatedness and inclusive fitness can be used in a dynamic model where frequencies of genotypes are changing over time; we use these calculations with an appropriate dynamics to see how the population will evolve and to find the equilibria. For this model, we use the selection-mutation dynamics, which is just like the replicator dynamics except that there is an extra term that keeps track of mutations.[14]

Figure 4 shows the dynamical analysis of this model, using both inclusive fitness and neighbor-modulated fitness. Figures 4(a) and 4(b) show, respectively, how the inclusive fitness and neighbor-modulated fitness change as the population composition changes. Figures 4(c) and 4(d) show the evolutionary trajectories in the population, in terms of the change in frequency of altruists. When this change is positive (when the pink line is above the x-axis, which is represented by the black dashed line in figures 4(c) and (d)), altruists will increase in frequency. Likewise when the change is negative, altruists will decrease in frequency. Information about the magnitude of selective pressures is also represented; the further the pink line is from zero, the more selective pressure there is and the faster the population composition will change.

[14] With the selection-mutation dynamics, a population with two types will evolve according to the following equation: $\dot{x}_t = x_t[f_t(x) - \bar{f}(x)] + \mu(1 - 2x_t)$. Note that since this is the same as the replicator dynamics except for the mutation term, which does not depend on the definition of $f_t(x)$, we can prove that using neighbor-modulated fitness and inclusive fitness will be equivalent in the same way as in section 4.1.

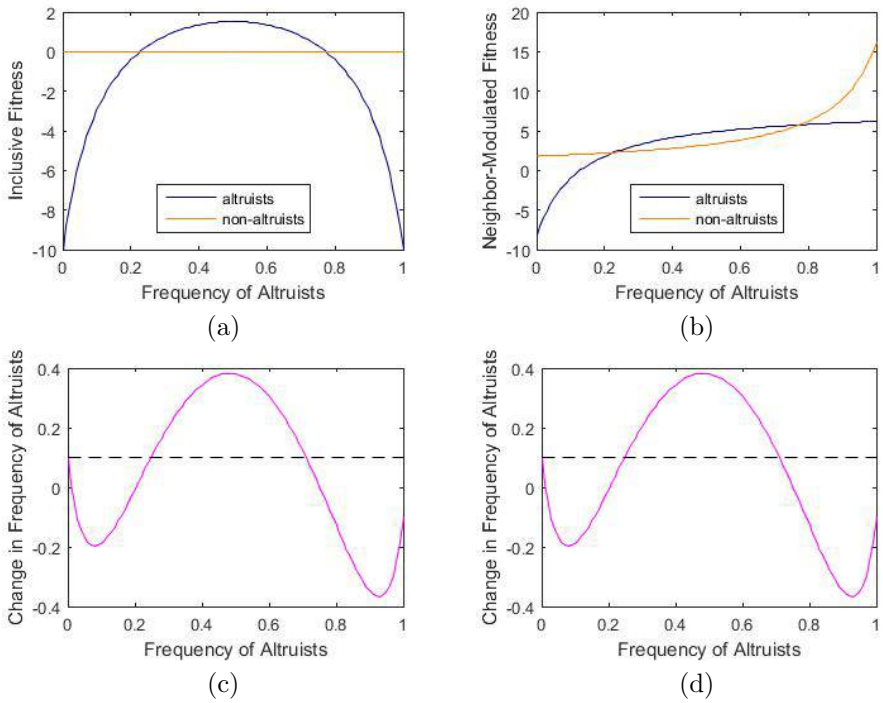

Figure 4: A comparison of inclusive fitness and neighbor-modulated fitness, for $\mu = 0.1$, $b = 18$ and $c = 10$. Comparing the calculations of inclusive fitness shown in (a) and neighbor-modulated fitness in (b) shows how the calculations of the two types of fitnesses differ. Comparing the change in the frequency of altruists found using inclusive fitness in (c) and neighbor-modulated fitness in (d) shows that the evolutionary trajectories are the same regardless of which calculation of fitness is used.

Comparing figures 4(a) and 4(b) shows that the two methods of calculating fitness do yield different numerical values of fitness. However, in comparing the evolutionary trajectory found using inclusive fitness in figure 4(c) with the trajectory calculated using neighbor-modulated fitness in figure 4(d), one can see that the choice between these fitness measures makes no difference for predicting the evolution of the population, either for the quantitative predictions of the amount of evolutionary change over time or the qualitative predictions about the evolutionary outcomes based on the model. That is, in this simple model, inclusive fitness and neighbor-modulated fitness both give us the same answer when we ask how much altruists will increase or decrease in frequency, across all possible population compositions.

We can also use either type of fitness calculation to find when the change in altruists is zero, when frequencies are not changing and the population is at an equilibrium. With the values of μ, b, and c chosen here, there are four equilibria, two of which are stable: one at about 1% altruists and one at about 75% altruists.[15]

5 Discussion

We can see from section 4 not only that inclusive fitness is perfectly well-suited for use in evolutionary game theory, but also that weak selection is not a necessary assumption for inclusive fitness calculations and that these calculations can be part of dynamically sufficient models. Some methods of calculating or estimating inclusive fitness may require stringent assumptions, but the calculations in general do not always require extra assumptions. How are we to understand this in the context of the debate over inclusive fitness?

5.1 Inclusive Fitness with Idealized Models

Some of the disagreement over inclusive fitness can be understood as arising from two sides of the debate emphasizing different methodologies. Recall from section 3.3 that inclusive fitness is seen as fundamentally within the quantitative genetics tradition, while critics of inclusive fitness tend to favor population genetics or evolutionary game theory. This means that inclusive fitness theorists tend to favor models which make use of abstractions, leaving details out while still providing literally true general claims about evolution. By contrast, evolutionary game theory, one of the preferred frameworks of the critics of inclusive fitness, tends to provide highly idealized models, making many assumptions which we know are not true of any real population but which allow us to develop a simple model of a fictional population that we think is similar to the real world in important ways.

As discussed in section 4.1, when there is an infinite population and a finite number of types, inclusive fitness calculations from quantitative genetics and

[15] An equilibrium is stable when selective pressures will cause the population to return to the equilibrium if a small amount of drift changes gene frequencies in the population.

evolutionary game theory are equivalent. Since quantitative methods are designed to handle continuously varying traits, assuming a finite number of types takes the methods out of the context in which they were developed and puts them into the context where dynamically sufficient models can be built. In doing so, we can get models with the kinds of properties valued by critics of inclusive fitness.

One way to think about this is that the regression methods developed by inclusive fitness theorists do not, in themselves, provide models with the properties the critics of inclusive fitness argue evolutionary models should have. However, we can formulate idealized models which are dynamically sufficient and which incorporate selection that is not weak. Then, when we abstract away from the particular details of genetic inheritance or population structure assumed by the simplified models, we arrive at the abstract equations based on the Price equation, which are often used in inclusive fitness theory. Section 4.1 (and the appendix) showed how, when we make simplifying assumptions commonly made in evolutionary game theory, the replicator dynamics and the versions of the Price equation often used in inclusive fitness theory are equivalent descriptions of evolutionary change.

Section 4.2 gives an example of how the regression methods commonly used in inclusive fitness can be seen as abstract descriptions of models within evolutionary game theory. In this simple model, we can track how $\frac{Cov(p,g')}{Cov(p,g)}$ changes as the population evolves. Using covariances might seem a bit unnatural in this overly simplified case: because we can calculate the relatedness directly from the assumptions of the model, we do not need to estimate it using the methods of quantitative genetics.

In fact, one might wonder whether there is any benefit to be gained from inclusive fitness in this sort of simplified model. One of the main perceived benefits of inclusive fitness is that it allows modelers to track changes in traits rather than the genes encoding for these traits (which are very difficult to discover empirically) while accounting for genetics by using relatedness (which is often not too difficult to estimate in real populations) (Queller, 1992). Because we abstract away from the mechanisms of genetic inheritance and how the genes encode for the trait of interest, summarizing this with a 'relatedness' parameter, we can develop a phenotypic model that still incorporates genetics in a way that can be empirically easy to measure. That is, one can account for genetics without knowing or making assumptions about the actual underlying genetics of a trait. When we switch to an evolutionary game theoretic or a population genetics model, like the replicator dynamics, we generally then must make assumptions about what these underlying genetics are. We no longer use relatedness to estimate genetic assortment; we can calculate the level of assortment directly.

So, to a certain extent one might think that it is appropriate that the debate over inclusive fitness is a debate over methods: although we can use inclusive fitness in the highly simplified models of evolutionary game theory, in doing so we lose some of the main benefits of the inclusive fitness framework. The

statistical methods used in inclusive fitness make the framework particularly useful, although these methods may require weak selection to split fitness effects into additive components and do not provide dynamically sufficient models. Further, if we view the debate as being about the methods commonly used in quantitative genetics, we can see where these criticisms come from.[16] That is, since inclusive fitness has often been seen as fundamentally quantitative, and since one of the main benefits of inclusive fitness (incorporating genetics in a way that is easy to estimate in real populations) is generally tied up in the statistical methods arising out of quantitative genetics, it makes sense that the debate over inclusive fitness will be in part a debate over methods. However, the status of inclusive fitness should not be decided by a debate over the use of methods for which inclusive fitness is seen as particularly beneficial.

The model in section 4.2 demonstrates that there can still be some benefit to calculating inclusive fitness rather than neighbor-modulated fitness even in models which are highly idealized, where the level of assortment can be calculated directly. The explanation given for why the population does not evolve to a population composed entirely of altruists was that relatedness drops off as the population becomes more uniformly altruistic. This sort of intuitive explanation is not readily available when using neighbor-modulated fitness. Because the terms describing how the benefits of altruism fall differentially on altruists are split between two different fitness calculations (one for the fitness of altruists and one for fitness of non-altruists), there is no parameter which systemically changes as the population composition changes that we can point to in order to explain why the fitness of altruists drops off as the population becomes more uniform.[17]

5.2 The Use of Hamilton's Rule

Hamilton's rule, the most famous result arising out of inclusive fitness theory, has been criticized for not being generally true, for not having any predictive power, and for being misleading in the absence of a particular model (Nowak et al., 2010). There is some truth to these claims. To get Hamilton's rule in the form $bR-c>0$, where c and b are interpreted as costs and benefits as described in section 2.2, one has to assume additive fitness components as we have been doing throughout this paper. If fitness components are not additive, then the rule will not give a correct description of a condition for the spread of a trait. Additionally, if we only have enough information to estimate b, c, and R at a particular point in time, we cannot predict the evolutionary outcome. Further, if $bR-c>0$ when we estimate these parameters, we might even be mislead into

[16] This is, of course, not to say that these are the only methods used in inclusive fitness theory, but that the critiques of inclusive fitness are often wrapped up in critiques of the statistical methods (see Allen et al., 2013, for instance).

[17] Others argue that inclusive fitness is valuable because it allows us to maintain the analogy of organisms acting as if they are maximizing fitness (Grafen, 2007b; West and Gardner, 2013; Okasha et al., 2014; Okasha and Martens, 2016), or more modestly that it allows us to explain the selection of social traits due to their casual contributions to fitness (Birch, 2016). These benefits would hold regardless of the methods one uses, and so are not addressed here.

thinking that the population will eventually be entirely altruistic if we forget that the value of any of these parameters can change as the population evolves. However, inclusive fitness theorists will generally agree to this (see Marshall, 2015, for example), but maintain that Hamilton's rule has both predictive and explanatory power. It is not immediately clear where the disagreement lies.[18]

The distinction between idealizations and abstractions can again be helpful in understanding part of the dispute. In particular, why should we expect Hamilton's rule to be true in general? Results derived within population genetics and evolutionary game theory are never true in general, as they rely on idealizations to achieve their simplicity. By contrast, Hamilton's rule is seen as a general result that is applicable to any real population one might wish to study. This fits well with its prominent role in quantitative genetics, relying on abstractions rather than idealizing assumptions to help provide "the general principles of social evolution theory" (Marshall, 2015, p. xiv).

In this vein, there is emphasis on providing a version of Hamilton's rule that is generally true. Hamilton's rule can be given in a very general form in which we do not have to assume any particular population structure or additive payoff affects (Gardner et al., 2011). Birch (2014b) and Birch and Okasha (2015) describe this in detail, but we can think of the 'cost' and 'benefit' terms in the rule as statistical associations between an organism's fitness and its own genotype (a self-effect) and its social partner's genotype (an other-effect), respectively. This general version of Hamilton's rule is true of any population. "In effect, this is because we are abstracting away from the complex causal details of social interaction to focus on the overarching statistical relationship between genotype and fitness" (Birch and Okasha, 2015, p. 24).

The question is then whether this version of Hamilton's rule has any predictive power. It can have predictive power if its components can be understood causally instead of just statistically. That is, if the self-effect and other-effect terms can be interpreted as ways in which the focal organism causally contributes to its own and its social partner's fitness, we have a model that can be used to make predictions rather than just a statistical summary of evolution within a population. However, as Birch and Okasha (2015) explain, it is not entirely clear when a causal interpretation can be provided.

There are, however, a variety of different rules that go under the name 'Hamilton's rule', each of which follows from different assumptions about the evolutionary process. We can describe these versions of Hamilton's rule as falling into three categories. There are 'special' versions of the rule (where the b and c terms are interpreted as payoffs in a model) and 'approximate' versions (which provide marginal approximations of the general versions of the rule) in addition to the 'general' version described above (Birch and Okasha, 2015).

In the version of Hamilton's rule in section 3.1, the b and c terms are interpreted as payoff from a game, or parameters in the model, so this can be thought of as a special version of Hamilton's rule. The fact that we derived

[18] See (Marshall, 2015, chapter 6, note 9) for an example of an inclusive fitness model where parameters can change as the population evolves.

a condition $bR - c > 0$ for the spread of altruism depends on the particular payoff structure of the model. If there were non-additive payoffs, we would have derived a different condition for the spread of altruism. Section 4 (and the appendix) illustrated how these general versions of Hamilton's rule describe the special versions from particular models. As mentioned in sections 3.1 and 3.3, there are also approximate versions of Hamilton's rule that require the assumption of weak selection to calculate relatedness or in order to split fitness effects into additive components. Thus, these rules abstract away from the particular payoff structure and so describe a wider range of cases than special forms of the rule. The assumption of weak selection, then, provides some restriction on the conditions under which approximate versions of Hamilton's rule will apply, but allows us to give an approximately correct condition for the spread of a social behavior for arbitrary payoff structures. (See Birch and Okasha (2015) for more discussion.)

Note that both general and approximate versions of Hamilton's rule apply for arbitrary payoff structures, but neither are dynamically sufficient. They instead allow us to perform a static analysis, comparing fitnesses at specific points in the evolutionary process (usually the points of interest are equilibria). Since this paper has looked at how inclusive fitness is used in the replicator dynamics compared with approaches based on the Price equation, it has focused on the the contrast between abstract models in quantitative genetics and idealized models in evolutionary game theory. However, that the critics of inclusive fitness prefer dynamic models over these static modeling techniques is perhaps the more fundamental disagreement in the debate.

There is the additional issue of interpreting the R parameter in Hamilton's rule. Although many inclusive fitness theorists recognize that R in inclusive fitness calculations can be thought of as a general measure of correlation, Hamilton's rule is still usually presented as a condition for the evolution of a trait by kin selection. However, this is an additional opportunity for Hamilton's rule to be misleading; a suggested biological or causal interpretation of the parameter might be unwarranted. Some criticisms seem to assume that Hamilton's rule is only useful when R is a measure of kinship (Nowak et al., 2010). The thought behind these sort of critiques of Hamilton's rule seems to be that when R does not have an intuitive biological interpretation, it is not clear what explanatory power is gained from forcing terms into this particular inequality. The power of Hamilton's rule then comes from using something like the statistical definitions of relatedness provided here and estimating relatedness using measures of kinship, like pedigrees.[19]

Since the statistical definitions of relatedness are historically explained and used as measures of kinship, adopting Hamilton's rule as a starting point might seem to suggest an interpretation in terms of kin selection and may lead to

[19]There are of course, other issues with applications of Hamilton's rule aside from interpreting R in terms of kinship. Often in more biologically realistic models, in order to keep R defined in a way that is plausibly connected to relatedness, b and c become functions of R itself. These sorts of issues are dealt with by Frank (2013); Birch and Okasha (2015) among others.

theorists ignoring other mechanisms that generate assortment between types. Connecting the statistical and probabilistic definitions here is one way of emphasizing how the association between 'relatedness' and R is contingent: R just measures differences in conditional probabilities of interacting with certain types in the population. In this context, Hamilton's rule might be thought of as a convenient mathematical description of the fact that there must be sufficient assortment between types in order for a trait such as altruism to evolve, a general point that has been made without the use of Hamilton's rule (see Skyrms, 1996, for example). A fully specified (but idealized) model, like the one in section 4.2, can connect R in Hamilton's rule to kinship, giving it a meaningful biological interpretation.

This is in line with one suggestion to avoid wrongly interpreting results in terms of kin selection, advanced by Taylor and Frank (1996) and Frank (2013), among others: formulate and analyze a model first, then afterwards use Hamilton's rule to give an intuitive explanation of the results if appropriate. This allows us to set up the model with whatever mechanism of assortment we think is plausible, then use Hamilton's rule if it helps illuminate important aspects of the causal structure.

6 Conclusion

While there can be benefits to using inclusive fitness, this does not mean that it is always beneficial to do so. Whether inclusive fitness or Hamilton's rule should be used depends on the model or the population one is studying. Many of the issues involved in deciding whether to use these methods were not addressed here. This paper has discussed the use of inclusive fitness in a special type of evolutionary model, in which pairwise interactions, additive fitness effects, and a finite number of types were assumed. In doing so, this paper focused the discussion on issues surrounding the different methodologies favored by the critics and proponents of inclusive fitness theory, in absence of conceptual and mathematical complexities that can arise in more complicated scenarios. Looking at this simple case helped to illuminate several features of the mathematical framework of inclusive fitness and the debate surrounding it.

While there may be difficulties with partitioning fitness effects into the form demanded by inclusive fitness when interactions become more complicated, we have seen that the specific causal partition used in inclusive fitness does not prevent one from building dynamically sufficient models nor does it require weak selection. Criticisms of inclusive fitness claiming that it requires these stringent assumptions are best thought of as criticisms of the types of quantitative methods generally used by inclusive fitness theorists. One can use inclusive fitness calculations in the sort of population genetic or evolutionary game theoretic models favored by these critics. In these models much of the advantage of using inclusive fitness, such as providing terms that can be easy to estimate empirically, disappears, but its power as an intuitive explanation of the evolution of social traits remains.

A Equivalence with Neighbor-Modulated Fitness

A.1 The Price equation describes the replicator dynamics

Following the definition provided in section 2.1, we can calculate the neighbor-modulated fitness of a pairwise interaction as follows:

$$f_i = s_{ii} + s_{-ii} \tag{13}$$

Keeping track of probabilities of receiving payoffs was necessary in section 4.1 in order to show the connection between neighbor-modulated fitness and inclusive fitness, but since we are only dealing with neighbor-modulated fitness we can use this less complicated expression. In these calculations, we will track the change in g, genetic value.

By definition, $\dot{E}(g) = \sum_i g_i \dot{x}_i + \sum_i \dot{g}_i x_i$. As mentioned, for simplicity we will assume there is no transmission bias and set $\sum_i \dot{g}_i x_i = 0$. Then, since the replicator dynamics provides us an equation for $\dot{x}_i$, we can plug the replicator dynamics into the Price equation:

$$\begin{aligned}
\dot{E}(g) &= \sum_i g_i \dot{x}_i \\
&= \sum_i g_i x_i [s_{ii} - \frac{1}{n}\sum_j s_{jj} + s_{-ii} - \frac{1}{n}\sum_j s_{-jj}] \\
&= \sum_i g_i x_i s_{ii} - \sum_i g_i x_i \frac{1}{n}\sum_j s_{jj} + \sum_i g_i x_i s_{-ii} - \sum_i g_i x_i \frac{1}{n}\sum_j s_{-jj} \\
&= E(s_{ii}g) - E(s_{ii})E(g) + E(s_{-ii}g) - E(s_{-ii})E(g) \\
&= Cov(s_{ii}, g) + Cov(s_{-ii}, g)
\end{aligned} \tag{14}$$

This is the Price equation with fitness partitioned into two components, the effect the focal organism has on its own fitness and the effect the social partner has on the focal organism's fitness. Theorists often derive this from the original Price equation in order to use neighbor-modulated fitness calculations and introduce relatedness calculations (see Queller, 1992, for example).

Hamilton's rule can easily be derived from this equation. Since the way an organism affects the fitness of itself and others is (to a certain degree) predicted by its phenotype, we can write both fitness terms as the following regressions:

$$s_{ii} = \alpha_{s_{ii}p} + \beta_{s_{ii}p} \cdot p + \epsilon_{s_{ii}p} \tag{15}$$

$$s_{-ii} = \alpha_{s_{-ii}p'} + \beta_{s_{-ii}p'} \cdot p' + \epsilon_{s_{-ii}p'} \tag{16}$$

Since the α's are the intercepts of the regression, they are constants and cannot covary with g. The ϵ's are error terms, which do not covary with g when payoffs

are additive (Queller, 1992). So, plugging (15) and (16) into (14), we're left with:

$$\dot{E}(g) = \beta_{s_{ii}p} Cov(p, g) + \beta_{s_{-ii}p'} Cov(p', g) \tag{17}$$

When we can interpret $\beta_{s_{ii}p}$ as a cost c and $\beta_{s_{-ii}p'}$ as a benefit b this gives us:

$$\dot{E}(g) > 0 \text{ when } b \cdot \frac{Cov(p', g)}{Cov(p, g)} - c > 0 \tag{18}$$

where $\frac{Cov(p',g)}{Cov(p,g)}$ is the neighbor-modulated fitness version of relatedness.

A.2 The replicator dynamics describes the Price equation

When there are a finite number of types, g_i can be written as an indicator function:

$$g_j^{<i>} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

For Page and Nowak (2002), who were considering phenotypes rather than genotypes, assuming a finite number of types was a restriction. Here, in considering genotypes, it is a natural assumption to make.

We can then use this indicator function in the Price equation with two fitness components derived above, and simplify:

$$\begin{aligned}
\dot{E}(g) &= Cov(s_{ii}, g^{<i>}) + Cov(s_{-ii}, g^{<i>}) \\
&= E(s_{ii} g^{<i>}) - E(s_{ii}) E(g^{<i>}) + E(s_{-ii} g^{<i>}) - E(s_{-ii}) E(g^{<i>}) \\
&= \sum_j g_j^{<i>} x_j s_{ii} - \sum_j g_j^{<i>} x_j \frac{1}{n} \sum_j s_{jj} + \sum_j g_j^{<i>} x_j s_{-ii} - \sum_j g_j^{<i>} x_j \frac{1}{n} \sum_j s_{-jj} \\
&= x_i s_{ii} - x_i \frac{1}{n} \sum_j s_{jj} + x_i s_{-ii} - x_i \frac{1}{n} \sum_j s_{-jj} \\
&= x_i [f_i(x) - \bar{f}]
\end{aligned} \tag{19}$$

Since $g_j^{<i>} = 1$ when $i = j$ and 0 otherwise, $\sum_j g_j^{<i>} x_j = x_i$, and this simplifies to yield the replicator dynamics.

B Equivalence with Inclusive Fitness

B.1 The Price equation describes the replicator dynamics

This is done in the same way as appendix A, except we take into account that the genetic value of the focal organism times its relatedness to its social partner

is a measure of the social partner's genetic value:

$$\begin{aligned}\dot{E}(g) &= \sum_i g_i(x_i[s_{ii} - \frac{1}{n}\sum_j s_{jj} + x_i r_{i-i}[s_{i-i} - \frac{1}{n}\sum_j s_{j-j}]) \\ &= \sum_i g_i x_i s_{ii} - \sum_i g_i x_i \frac{1}{n}\sum_j s_{jj} + \sum_i g_i r_{i-i} x_i s_{i-i} - \sum_i g_i r_{i-i} x_i \frac{1}{n}\sum_j s_{j-j} \\ &= Cov(s_{ii}, g) + Cov(s_{i-i}, g')\end{aligned} \tag{20}$$

This is again a version of the Price equation where the fitness effect is split into two components. Here, though, fitness is split into the effect the focal organism has on its own fitness and the effect the focal organism has on its social partner's fitness.

In order to relate this to Hamilton's rule, we can again notice that the fitness components are predicted by phenotype. Since in this case the focal organism causes the fitness effects, both for itself and its social partner, the phenotype of the focal organism predicts both fitness effects. So, we use the phenotype of the focal organism in both regressions:

$$s_{ii} = \alpha_{s_{ii}p} + \beta_{s_{ii}p} \cdot p + \epsilon_{s_{ii}p} \tag{21}$$

$$s_{-ii} = \alpha_{s_{i-ip}} + \beta_{s_{i-ip}} \cdot p + \epsilon_{s_{i-ip}} \tag{22}$$

We can then plug (21) and (22) into (20) and rearrange to obtain the inclusive fitness version of Hamilton's rule:

$$\dot{E}(g) > 0 \text{ when } b \cdot \frac{Cov(p, g')}{Cov(p, g)} - c > 0 \tag{23}$$

B.2 The replicator dynamics describes the Price equation

We can again let g_i be an indicator function and write:

$$\begin{aligned}\dot{E}(g) &= Cov(s_{ii}, g^{<i>}) + Cov(s_{-ii}, g'^{<i>}) \\ &= \sum_j g_j^{<i>} x_j s_{ii} - \sum_j g_j^{<i>} x_j \frac{1}{n}\sum_j s_{jj} + \sum_j g_j^{<i>} r_{j-j} x_j s_{i-i} - \sum_j g_j^{<i>} r_{j-j} x_j \frac{1}{n}\sum_j s_{j-j} \\ &= x_i s_{ii} - x_i \frac{1}{n}\sum_j s_{jj} + x_i r_{i-i} s_{i-i} - x_i r_{i-i} \frac{1}{n}\sum_j s_{j-j} \\ &= x_i[f_i(x) - \bar{f}]\end{aligned} \tag{24}$$

Again, this simplifies to yield the replicator dynamics.

References

Abbot, P., J. Abe, J. Alcock, S. Alizon, J. A. C. Alpedrinha, and et al. (2011). Inclusive fitness theory and eusociality. *Nature* *471*, E1–E4.

Allen, B. and M. A. Nowak (2015). Games among relatives revisited. *Journal of theoretical biology 378*, 103–116.

Allen, B., M. A. Nowak, and E. O. Wilson (2013). Limitations of inclusive fitness. *Proceedings of the National Academy of Sciences 110*, 20135–20139.

Birch, J. (2014a). Gene mobility and the concept of relatedness. *Biology & Philosophy 29*(4), 445–476.

Birch, J. (2014b). Hamilton's rule and its discontents. *British Journal for the Philosophy of Science 65*(2), 381–411.

Birch, J. (2016). Hamilton's two conceptions of social fitness. *Philosophy of Science 83*(5).

Birch, J. and S. Okasha (2015). Kin selection and its critics. *BioScience 65*(6), 22–32.

Chakraborty, A. and R. Harbaugh (2007). Comparative cheap talk. *Journal of Economic Theory 132*(1), 70–94.

Eliashberg, J. and R. L. Winkler (1981). Risk sharing and group decision making. *Management Science 27*(11), 1221–1235.

Frank, S. A. (1998). *Foundations of social evolution*. Princeton University Press.

Frank, S. A. (2013). Natural selection. vii. history and interpretation of kin selection theory. *Journal of Evolutionary Biology 26*, 1151–1184.

Gardner, A., S. West, and G. Wild (2011). The genetical theory of kin selection. *Journal of Evolutionary Biology 24*(5), 1020–1043.

Godfrey-Smith, P. (2009). *Abstractions, idealizations, and evolutionary biology*. Springer.

Grafen, A. (1979). The hawk-dove game played between relatives. *Animal behaviour 27*, 905–907.

Grafen, A. (1984). Natural selection, kin selection and group selection. In J. Krebs and N. Davies (Eds.), *Behavioural Ecology* (2 ed.). Oxford: Blackwell Scientific Publications.

Grafen, A. (1985). A geometric view of relatedness. *Oxford surveys in evolutionary biology 2*(2).

Grafen, A. (2007a). Detecting kin selection at work using inclusive fitness. *Proceedings of the Royal Society of London B: Biological Sciences 274*(1610), 713–719.

Grafen, A. (2007b). The formal darwinism project: a mid-term report. *Journal of evolutionary Biology 20*, 1243–1254.

Hamilton, W. D. (1963). The evolution of altruistic behavior. *The American Naturalist* *97*(896), 354–356.

Hamilton, W. D. (1964). The genetical evolution of social behavior i and ii. *Journal of Theoretical Biology* *7*, 1–16.

Hamilton, W. D. (1975). Innate social aptitudes of man: an approach from evolutionary genetics. *Biosocial anthropology* *133*, 155.

Hammerstein, P. and R. Selten (1994). Game theory and evolutionary biology. In S. Hart (Ed.), *Handbook of Game Theory with Economic Applications*, Volume 2. Amsterdam: Elsevier Science.

Lehmann, L. and F. Rousset (2014). The genetical theory of social behaviour. *Philosophical Transactions of the Royal Society B* *369*(1642).

Maciejewski, W., F. Fu, and C. Hauert (2014). Evolutionary game dynamics in populations with heterogenous structures. *PLoS Comput Biol* *10*(4), e1003567.

Marshall, J. A. (2011). Queller's rule ok: Comment on van veelen 'when inclusive fitness if right and when it can be wrong'. *Journal of Theoretical Biology* *270*, 185–188.

Marshall, J. A. (2015). *Social Evolution and Inclusive Fitness Theory*. Princeton University Press.

Nowak, M. A., C. E. Tarnita, and E. O. Wilson (2010). The evolution of eusociality. *Nature* *466*(26), 1057–1062.

Nowak, M. A., C. E. Tarnita, and E. O. Wilson (2011). Nowak el al. reply. *Nature* *471*, E9–E10.

Okasha, S. and J. Martens (2016). Hamilton's rule, inclusive fitness maximization, and the goal of individual behaviour in symmetric two-player games. *Journal of evolutionary biology*.

Okasha, S., J. A. Weymark, and W. Bossert (2014). Inclusive fitness maximization: An axiomatic approach. *Journal of theoretical biology* *350*, 24–31.

Orlove, M. and C. L. Wood (1978). Coefficients of relationship and coefficients of relatedness in kin selection: a covariance form for the rho formula. *Journal of Theoretical Biology* *73*(4), 679–686.

Page, K. M. and M. A. Nowak (2002). Unifying evolutionary dynamics. *Journal of Theoretical Biology* *270*, 93–98.

Queller, D. C. (1992). Quantitative genetics, inclusive fitness, and group selection. *The American Naturalist* *139*(3), 540–58.

Rousset, F. (2002). Inbreeding and relatedness coefficients: what do they measure? *Heredity* *88*, 371–380.

Skyrms, B. (1996). *Evolution of the Social Contract.* Cambridge University Press.

Skyrms, B. (2002). Altruism, inclusive fitness, and the logic of decision. *Philosophy of Science* *69*, S104–111.

Taylor, P. D. Wild, G. and A. Gardner (2007). Direct fitness or inclusive fitness: how shall we model kin selection? *Journal of Evolutionary Biology* *20*, 301–309.

Taylor, P. D. and S. A. Frank (1996). How to make a kin selection model. *Journal of Theoretical Biology* *180*, 26–37.

Taylor, P. D. and W. Maciejewski (2012). An inclusive fitness analysis of synergistic interactions in structured populations. *Proceedings of the Royal Society London B* *279*(1747).

Traulsen, A. (2010). Mathematics of kin- and group-selection: formally equivalent? *Evolution* *64*(2), 316–323.

van Veelen, M. (2005). On the use of the price equation. *Journal of Theoretical Biology* *237*, 412–426.

van Veelen, M. (2009). Group selection, kin selection, altruism and cooperation: when inclusive fitness if right and when it can be wrong. *Journal of Theoretical Biology* *259*, 589–600.

van Veelen, M. (2011). The replicator dynamics with n players and population structure. *Journal of theoretical biology* *276*(1), 78–85.

van Veelen, M., J. García, M. W. Sabelis, and M. Egas (2012). Group selection and inclusive fitness are not equivalent; the price equation vs. models and statistics. *Journal of theoretical biology* *299*, 64–80.

West, S. A. and A. Gardner (2013). Adaptation and inclusive fitness. *Current Biology* *23*(13), R577–R584.

Wild, G. and A. Traulsen (2007). The different limits of weak selection and the evolutionary dynamics of finite populations. *Journal of Theoretical Biology* *247*(2), 382–390.

Wilson, E. O. (2012). *The Social Conquest of Earth.* W. W. Norton.